Trainable Cataloging for Digital Image Libraries with Applications to Volcano Detection

M.C.Burl[†‡], U. M. Fayyad[‡], P. Perona[†], P. Smyth[‡]

t California Institute of Technology
MS 116-81 --- Pasadena, CA 91125
{burl,perona}@systems.caltech.edu

[‡] Jet Propulsion Laboratory, California Institute of Technology
MS 525-3660 -- Pasadena, CA 91109
{fayyad,pjs}@aig.jpl.nasa.gov

June 30, 1995

Abstract

Users of digital image libraries are often not interested in image data *per se* but in derived products such as catalogs of objects of interest. Converting an image database into a usable catalog is typically carried out manually at present. For many larger image databases the purely manual approach is completely impractical. In this paper we describe the development of a trainable cataloging system: the user indicates the location of the objects of interest for a number of training images and the system learns to detect and catalog these objects in the rest of the database. In par titular we describe the application of this system to the cataloging of small volcanoes in radar images of Venus. The volcano problem is of interest because of the scale (30,000 images, order of 1 million detectable volcanoes), technical difficulty (the variability of the volcanoes in appearance) and the scientific importance of the problem. The problem of uncertain or subjective ground truth is of fundamental importance in cataloging problems of this nature and is discussed in some detail. Experimental results are presented which quantify and compare the detect ion perfor mance of the system relative to human detection performance. The paper concludes by discussing the limitations of the proposed system and the lessons learned of genera] relevance to the development of digital image libraries.

Keywords: digital image libraries, pattern recognition, science data analysis, volcanoes, Venus, SAR, detection, classification, learning, remote sensing

1 Introduction

In recent years there have been significant advances in image acquisition and storage technologies. Large image dat abases in fields as diverse as astronomy, geology, and diagnostic medicine are increasingly routine. However, our ability to *analyze* data lags fat behind our ability to *collect* data. Users of image databases, such as astronomers, geologists, and medical experts, are not interested in the in iage data *per* se. The images are but an intermediate representation from which hypotheses can be inferred about the physical properties of the targets being imaged. Many image database users wish to work with derived image products, such as catalogs of objects of interest.

For example, in planetary science, the scientific process involves examination of images (and other data) from planetary bodies such as Venus and Mars, the conversion of these images into catalogs of geologic objects of interest, and the use of these catalogs to support, refute, or originate t heories about t he geologic evolu t ion and current state of the planet. Typically these catalogs contain information abou t the location, size, shape, and general context of the object of interest and arc published and made generally available to the planetary science community [1].

In the last 3(I years remote spacecraft have provided far more detailed planetary images than were previously available, and subsequently our understanding of the physical geology of the planets has increased substantially. Nonetheless much remains to be discovered and the scientific process is ongoing. Traditionally the analysis of planetary surface images has been a manual process where much of the work was carried out by geologists analyzing hard-copy images. There is currently a significant shift to computer-aiclcd processing of planetary data, a shift which is driven by the public availability of many planetary datasets in digital form on CD-ROMS [2]. Thus far, however, the geologist's routine remains largely manual: the computer is used as a storage and display tool, but is hardly used for automated analysis. Hence, what could potentially be turned into a large digital image library is simply stored as raw image data. Given the volume of data being collected (see for example Section 3) purely manual cataloging of objects of interest is completely impractical. Thus, as in the volcano problem discussed later, scientists are manually cataloging small portions of the dataset and inferring what they can from these data [3].

In this paper wc describe a system for automatically locating small volcanoes on the surface of Venus. The cataloging; and study of volcanoes on Venus is itself an important scientific problem, yet it can also be considered a typical instance of a common problem in image database exploration: a user can identify a number of examples of an object of interest and would like the system to automatically find and characterize all such objects in the image database. Our approach relies heavily on the notion that the system is trainable and can learn a detection model from the identified examples. We view this as being far preferable to the primary alternative which would be to program a special-purpose model for each object. With the trainable system approach, a user can modify the detection model at will in an interactive manner by identifying specific training examples of interest in a natural manner. Thus, the benefits are clear. Whether the trainable approach is technically feasible is not clear: we will return to this issue in Section 8.

The main issues discussed in this paper are:

- The collection and handling of training data from the users
- The implicit subjective nature of image labelling by human experts,
- The evaluation and comparison of human and algorithm performance in the absence of absolute ground truth.
- The technical issues of detection, feature extraction, and classification which are critical to designing trainable cataloging systems.

Each of t hcsc issues is cl irect ly relevant to the problem of creating digital image libraries from raw image data. In particular, we focus on the problem of converting, original image data into digital catalogs which provide a high-level link to the original data for access and exploration.

The paper begins by discussing relevant prior work on the detection of natural objects in remote-scrlsing imagery (Section 2). Section 3 describes the Magellan mission to the planet Venus and provides more motivation and detail about the volcano detection problem. The nature of the Magellan images is discussed in Section 4. Section 5 describes how the training data is generated and focuses in particular on the problems associated with not having absolute ground truth. In Section 6 a three-stage volcano detection system is introduced and described in some detail. Experimental results are present ed in Sect ion 7, where the volcano detection system is quantitatively evaluated with respect t o human det ection performance on several test sets of Magellan images. Finally, Section 8 discusses t he lessons learned from this project and their relevance to more general digital image library prol)lems.

2 Prior Work on Detecting Natural Objects in Remote-Sensing Data

Prior work using patt ern recogni tion wit h remote sensing data has largely focused on earth-based data and the classification of homogeneous regions int o vegetation types (for example) [4]. Most work on the detection of ol)jects in remotely-sensed data has largely been limited to the detection of man-made ol)jects wit h well-defined edge characteristics. Indeed in an overall sense there is little prior work on the detect ion of multiple natural objects in a noisy environment - many techniques implicitly assume that t he object of interest has already been located in the image and focus on the problem of finding good discriminants to compare object hypotheses. Hough transform methods have been used in the past for detection of circular geologic features in SAR data [5, 6] but without great success. In the particular context of the volcano detection problem, Wiles and Forshaw [7] described a matched filtering approach for detection of small volcanoes in the Magellan data, In Section 7 we will see that matched filtering alone appears insufficient to achieve high detection rates for this problem. Note also that these methods involve relatively little, if any, training based on expert-sulq)lic(l data. In contrast, the approach proposed here emphasizes the notion of a trainable system which t he user can customize at will by providing specific examples of the object to be detected.

Problems with many similar characteristics to the volcano problem occur in medical diagnostic imaging, for example automated analysis of tissue abnormalities in pathology or detection of tumors in magnetic resonance imaging. In general t hcsc methods take great advantage of the fact that they are in ^ controlled environment and, hence, can usc a clearly contrasting background with reference points. This leads to a much higher effective signal-to-noise ratio than one encounters in the Venus volcano images.

a ^

3 Venus volcanism

3.1 Background on the Magellan Mission to Venus

May 4, 1989

On May 4th, 1989, the Magellan spacecraft was launched from Earth on a mapping mission to Venus. Magellan entered an elliptical orbit around Venus in August 1990 and subsequently transmitted back to Earth more data than that from all past planetary missions combined [8]. In particular, a set of approximately 30,000, 1024 x 1024 pixel, synthetic aperture radar S(AR), 75m/pixel A" resolution images of the planet's surface were transmitted, r esulting in a high resolution map of 97% of the surface of Venus. The total combined volume' of pre-Magellan image data available from various past US and USSR spacecraft and groulld-based observations represents only a tiny fraction of the Magellan data set. Thus, the Magellan mission has provided planetary scientists with an unprecedented data set for Venus science analysis. It is anticipated that the study of the Magellan dataset will continue well into the next century [1, 9, 10].

Tile Magellan image dab+'t is a unique testbed for prototyping digital image library tools: it is of significant scientific importance, it is large enough that automated and semi-automated tools arc essential if even a fraction of the data is to be utilized, it has an enthusiastic user community (planetary geologists) who are ready to use t hcse tools, and it contains significant technical challenges in t erms of pattern recognition and image an alysis (as we shall see in more detail in this paper). All the scientific data from t he mission has been publicly released by NASA in digital form on CD-R OMS ensuring widespread low-cost access.]n Appendix 1 wc describe how the data used in the experiments in this paper can be obtained.

3.2 The Scientific Importance of Venus Volcanism

The location, identification, and cataloging of volcanoes are key components in the study of Venus. To quote Saunders et al [8]:

> Volcanism is the most widespread and important geologic phenomenon on Venus. Volcanic features are broadl y distributed globally, unlike plate boundary concentrations typical of Earth. The most widespread t errain type on Venus is lowland volcanic plains.

Understanding clustering characteristics and the global distribution of the volcanoes is fundamental to understanding the regional and global geologic evolution of the planet [3], Generating a comprehensive catalog including the size, location, and other relevant informat ion about each volcano is clearly a pre-requisite for more advanced studies sucl I as cluster analysis of the volcano locations, This catalog can potentially provide the data necessary to answer basic questions concerning the geophysics of Venus, which is of' particular interest since geologically, Venus is Earth's sister planet. Typical geophysical questions about Venus volcanism concern eruption mechanics, the relationship between volcanoes and local t ectonic structure, and the pattern of heat flow within the planet.

Geologists estimate the number of small volcanoes (diameter < 15km) on the planet to be $\sim 10^6$ [II]. These volcanoes are t hought to be widely scattered throughout 30,000 1Mbyte images. Manually locating these volcanoes is simply not feasible. We have typically found in our experiments that humans t end to fat igue quickly after labelling on the order of 50 or 100 images over a time-scale of a few days. Thus, large-scale sustained cataloging by geologists is not realistic even if they could devote the necessary time to this task. In this context, an automated system for the detection and cataloging of volcanoes has considerable ut ility. From a more general digital library perspective we arc targeting the automat ion of the expensive and onerous cataloging step which is necessary to turn to a collection of images into an indexed and accessible digital library.

4 Magellan Imagery

A fundamental objective of the Magellan mission was to provide global mapping of the surface of Venus. The mapping was performed using synthetic aperture radar (SAR) because of its ability to penetrate the dense cloud cover surrounding Venus. The wavelength of the radar was 12.6 cm corresponding to a frequency of 2.385 GHz. The incidence angle varied from 15° to 45° and the number of looks varied from 5 to 16. Because the number of looks is relatively high this results in an effective averaging of the speckle noise which is commonly observed in SAR images: consequently the noise in the Magellan images is closer to the standard additive white noise typical of optical imaging. A complete description of the Magellan SAR imaging system is given in [12].

A standard Magellan image consists of 1024 x 1024 8-bit pixels, where the pixels are 75m in resolution for the results referred to in this paper, S1 nail volcano diameters are typically in the 2- 3km range, i.e., 30 to 50 pixels wide. Vole.anocs are often spatially clustered in volcano fields. As a consequence, most of the volcanoes are expected to be found in about 10-20% of the total number of images, and within these images there may number as many as 100 or more volcanoes, although typically the number is in the 10-50 range.

Figure 1 shows a 30km x 30km area imaged by Magellan (illumination from the left). This area located near (lat 30°N, lon 332°) contains many small volcanoes. Observe that the larger volcanoes in this figure have the classic radar signature one would expect based on the topography; that is, the upward sloping surface of the volcano in near-range (close to the radar) scatters more energy back to the sensor than the surrounding flat plains and therefore appears bright. The downward sloping surface of t he volcano in far-range scat, ters energy away from the sensor and therefore appears dark, Together, these effects cause t he volcano to appear as a left-twright *bright-dark pair* within a circular planimetric outline. Near the center of the volcanoes, there is usually a summit pit that appears as a *dark-bright* pair *because* t he radar energy backseat ters strongly from the far-range rim. Small pits, however, may not appear or may appear as only a bright spot due to the image resolution.

The topography-induced features described above are the primary visual cues that geologists report using to locate volcanoes. 1 low'ever, there are a number of other, more subtle cues. The apparent brightness of' an area in a radar image depends not only on the macroscopic topography but also on the surface roughness relative to the radar wavelength. Thus, if the flanks of a volcano

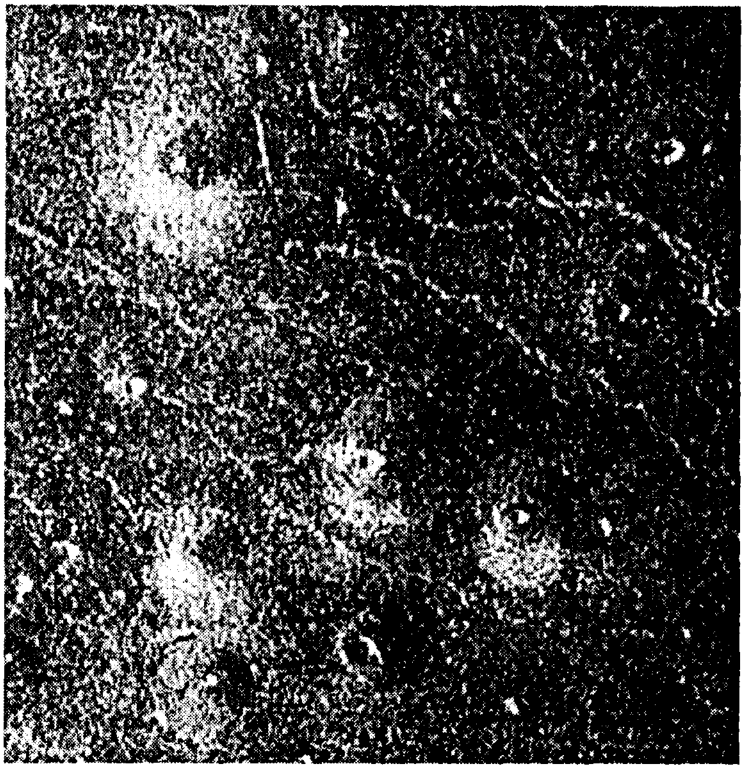

Figure 1: Magellan SAR sill>-ilna~c: A 30km x 30km region containing a number of small volcanoes. Illumination is from the left; incidence angle ≈ 40°.

have different roughness properties than the surrounding plains, the volcano may appear as a bright or dark circular area instead of as a bright-dark pair. Volcanoes may also appear as radial flow patterns, texture differences, or disruptions of graben. (Graben are ridges or grooves in the planet surface, which appear as bright lines in the radar imagery -- see Figure 1.)

5 obtaining a Labeled Training Database

In the volcano-location problem, as in many remote sensing applications, validated ground truth data does not exist. Due to the surface temperature of 482°C no remote landers have visited the surface of Venus apart from a Russian robotic lander which melted within a few minutes. Despite the fact that the Magellan data is the best imagery ever obtained of Venus, geologists cannot always determine with 100% certainty whether a particular image feature is indeed a volcano. This inherent ambiguity is due to factors such as image resolution, signal-to-noise level, interpretation of the SAR imagery, and so forth.

Sample Volcanoes

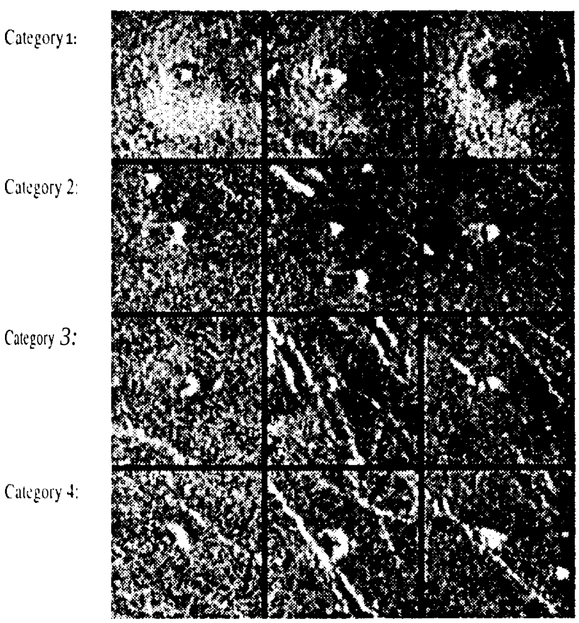

Figure 2: A selection of local regions as labeled by the geologists and their respective categories.

5.1 Volcano Categories

'1'here is considerable subjective variability in volcano labelling: for the same image, different geologists produce different label lists, and even the same geologist produces different lists over time. To help quantify this uncertainty, the geologists label training examples into quantized probability bins or "categories, " where the probability bins correspond to visually distinguishable sub-categories of volcanoes. In particular, 5 categories arc used:

1. where a summit pit, a bright-dark radar pattern, and apparent topographic slope are all clearly visible, probability 0.98,
2. where only 2 of the 3 criteria in category 1 arc visible, probability 0.80,
3. where no summit pit is visible but there is evidence of franks or a circular outline, probability 0.60,
4. where only a summit pit is visible, probability 0.50, and
5. where no volcauo-like features arc visible, probability 0.0.

The probability for category *i* corresponds to the incan probability that a volcano exists at a particular location *given* that it belongs to category i. These are *subjective* probability estimates and were elicited based on lengthy discussions with the planetary geologists. On average 10%, 20%, 40% and 30% of the volcanoes belong to the categories 1, 2, 3, and 4, respectively.

Figure 2 shows some typical volcanoes from each category. The use of quantized probability bins to attach levels of certainty to subjective image labels is not new: the same approach is routinely used in the evaluation of radiographic image displays to generate subjective ROC (receiver operating characteristic) curves [13, 14]. An ROC is useful for diagnostic applications since it displays the full range of possible operating thresholds for a detector (human or algorithmic). In contrast, the more oft-quoted probability of classification error criterion only represents a single point on the curve (typically the point at which the threshold on the posterior probability for deciding in favor of class 1 of each class is set to 0.5).

5.2 Handling Lack of Ground Truth

In the absence of absolute ground truth, the goal of a detection system is to be as comparable in performance as possible to the geologist's in terms of labelling accuracy. Absolute accuracy is not measurable for this problem. Hence, the best an automated detection system can do is to emulate the geologist's performance - this point will become clearer when performance metrics arc discussed later in the paper. Thus, unlike most supervised classification problems, the class labels are in fact subjective estimates of the true class labels as provided by experts (in this case, planetary geologists). This introduces two important issues to the traditional supervised classification problem.

5.2.1 Class Label Uncertainty During Training

The first issue concerns the training phase of the supervised learning algorithm. Since the measured class labels are uncertain estimates of the true class labels one can take this into account during training using a standard statistical decision theory approach. It can be shown that this amounts to weighting the examples according to posterior class membership probabilities [15]: if an example has probability 0.8 of belonging to class ω_1 and probability 0.2 of belonging to class W_2 then the example can be fractionally assigned during training to each class according to these weights. In a logistic regression context one can use the post crier probability probabilities directly as the target values. The practical issue is that of determining the posterior class probabilities: humans are notoriously poor at providing accurate estimates of subjective probabilities [16]. In the case where there is one set of labels per image, one approach is to map the expert's categories directly to posterior probability values as described above. Preliminary results indicate that the weighted approach provides no discernible performance improvement over tile non-weighted approach but these results were based on relatively small data sets [17],. For the case of multiple experts there are a variety of techniques available in the statistical literature for combining multiple expert ratings. We have also explored a relatively simple probabilistic model which results in a composite estimate based on the labels from different experts [18]. In this paper we only usc the relatively simple non-weighted method for training.

5.2.2 Performance Evaluation and Class Label Uncertainty

The second primary issue raised by class label uncertainty is that of evaluating relative performance of both humans and algorithms. If onc dots not know what tile absolute ground truth is, how can one evaluate the performance 01 any detector (be it human or algorithmic)? The answer is that while one cannot in general evaluate *absolute* detection performance, one can evaluate *relative* detection performance. The general approach we have taken is to evaluate the performance of algorithms and human experts against a reference labeling provided by another expert or set of experts. For example with two experts, one can compare both the algorithm and expert A relative to tile labeling of expert B, or B and the algorithm relative to A, or A and B individually and the algorithm relative to the consensus (joint) labelling by A and B. Once again one can use the estimated reference class probabilities to weight the performance criteria: if a detector classifies a local region as a volcano and the region has a probability of 0.6 (according to the reference data) of being a volcano, then onc could weight the performance criterion accordingly, e.g., the loss function would use the weight 0.6 in evaluating the performance [1 9}. In this paper we will adopt the simpler non-weighted method of performance evaluation just as we will use the non-weighted training classification t raining methods. We will sce later that the methodology of choice for evaluating relative performance involves variations of the receiver operating characteristic (the ROC).

5.3 Methodologies for Collecting Subjective Label Information

Participating in the development of the detection algorithm are planetary geologists from the Department of Geological Sciences, Brown University. We arc fortunate to have direct collaboration with two members of this group who were also members of the Volcanism Working group on the Magellan Science team. We will refer to these geologists as geologist A and B henceforth in this paper. Both of these geologists have extensive experience in studying both Earth-based and planetary volcanism and have published some of the standard reference works on Venus volcanism

[3, 1 1]. Hence, their collective subjective opinion is (roughly speaking) about as expert as one can find given the available data and our current state of knowledge about the planet Venus.

The standard manner in which we obtain labels is to have a labeller interact with an X-windows software too] whereby he or she uses mouse-clicks to locate candidate volcanoes. Starting with an initially blank image, the labeller proceeds to sequentially click on the estimated centers of the volcanoes. The labeller is then prompted to provide a subjective label estimate from a choice of categories 1- 4 as described earlier by default, locations which are not labelled are considered to have label "5" (non-volcano). Clearly it is possible that based on the visual evidence, for the same local image patch, the same label may not be pı ovided by different labellers, or indeed by the same labeller at different times. In addition to labels, the labeller can also provide a fitted diameter estimate by fitting a circle to the feature. Figure 3 shows the result of one such labelHing.

After completing the labelling, the result is an annotation of that image which can be stored in standard database format the unique key to the image is a label event, which corresponds to a particular latitude/longitude (to the resolution of the pixels) for a particular labeller at a particular time (since the same labeller may relabel an image multiple times). It is this database which provides the basic reference framework for deriving est imates of geologic parameters, training data for t he learn ing algorith ns, and reference data for per formance evaluation. A simple form of spatial clustering is used to determine which label events (from different labellers) actually correspond to the same geologic feature (volcano). It is fortunate that volcanoes tend not to overlap each other spatially and thus maintain a separation of at least a few kilometers, and also that different geologists tend to be quite consistent in their centring of the mouse-clicks --- mean differences of about 2.5 pixels (Euclidean distance) have been found in cross comparisons of label data from geologists A and 13, which is reasonable considering the precision one can expect from mouse location on a screen. 1 Ience, accurat c location of the volcanoes is not in itself much of problem. Figure 3 shows the results of a typical labeling session with a geologist.

5.4 Volcano Detection Performance of Human Experts

~'able 1 shows the confusion matrix between the two geologists for a set of 4 images. The (i, j)th

Table 1: Confusion Matrix of geologist A Vs. geologist B.

	geologist A				
	Label 1	Label 2	Label 3	Label 4	Not Detected
geologist B					
Label 1	19	8	4	1	3
Label 2	9	8	6	5	5
Label 3	13	12	18	1	37
Label 4	1	4	5	24	15
Not Detected	4	8	29	16	0

element of the confusion matrix counts the number of label events which correspond to labeller B generating label i and labeller A generating label j, where both labels were considered to belong to the same visual feature, i.e., were within a few pixels of each other. The (2, 5) (or (5, j)) entries count the instances where labeller B (or A) provided label i (or j), but labeller A (or B) did not provide any label entry (5, 5) is defined to be zero. Ideally, the confusion matrix would have all of its entries on the diagonal if both labellers agreed (completely on all events. Clearly, however, there is substantial disagreement, as judged by the number of off-diagonal counts in the matrix, For example, label 3's are parti cularly noisy, in both "directions." Label 3's are noisier than label 4's because there is less variability in the appearance of 4's compared to 3's (4's are simple pits, 3's are less well-defined). About 50% of the label 3's detected by either labeller are not detected at all by the other labeller. On the other hand, only about 10% of the! label 1's of either labeller are missed by tile other. This matrix underlines the inherent ambiguity present in this problem

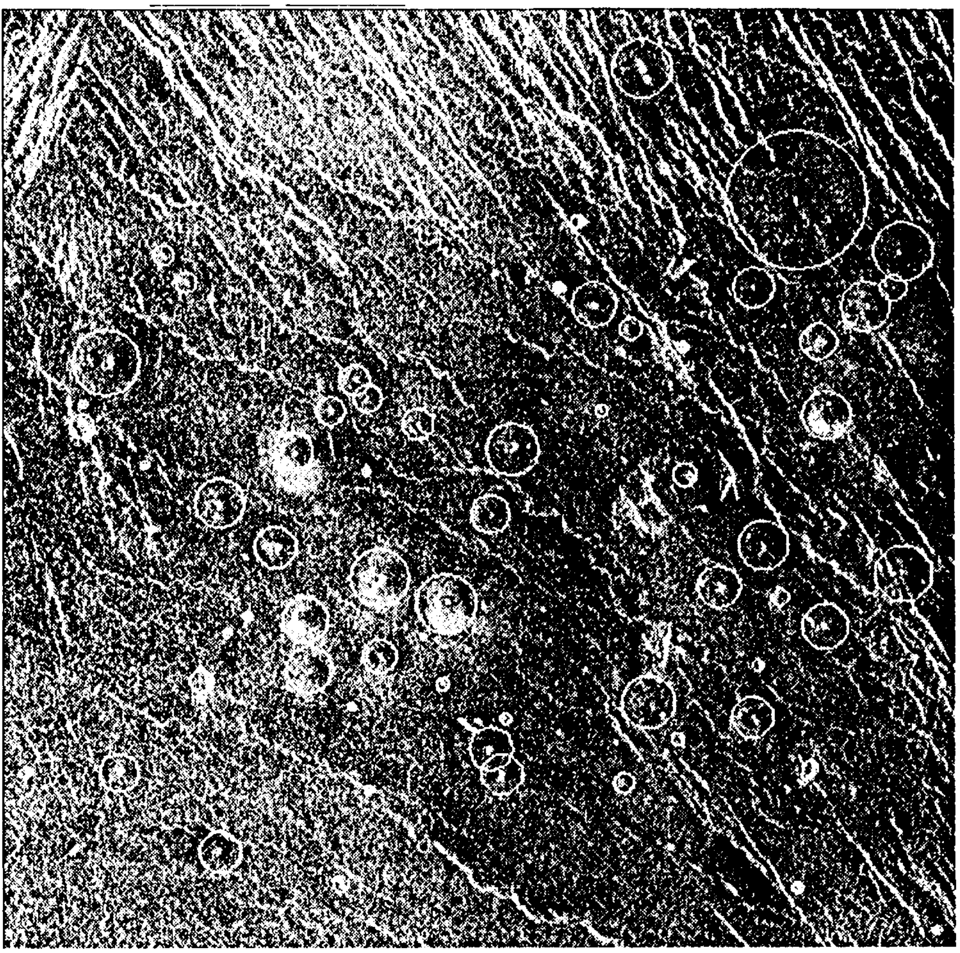

Figure 3: Magellan SAR image of Venus after labelling by a geologist showing estimated size and locations of small volcanoes.

even among experts. Thus, it is important to keep in mind that success for this task can only be measured in a relative manner: treating onc set of labels as ground truth, one measures how well the algorithmic detector compares to a human expel't in detection performance.

6 Description of the Volcano Detection Algorithm

In this section, we provide an overview of the algorithm we have developed for finding small volcanoes on Venus. We have decomposed the volcano detection problem into three sub-problems:

- *Detection* of candidate volcano regions in the original image
- *Feature extraction* from the detected local regions
- *Classification* of the extracted local features into volcano and non-volcano classes.

The transformation from high-dimensional pixel space to a lower dimensional feature space achieved by the feature extraction step is essential given that each volcano can typically occupy hundreds of pixels combined with the fact that relatively few positive training examples (order of hundreds) are available. Thus, direct use of the pixels as input to a classification algorithm is not practical given the ratio of input dimensionality to the number of training examples, Experimental results with a variety of feedforward neural network classification] models verified this hypothesis [20]. The training data were often linearly seprable in pixel space, resulting in an underconstrained training procedure whei e the model could memorize the ti aining data perfectly but generalized poorly to unseen data. separable

The detection step, which localizes the detector or focuses the attention of the detection algorithm on a local region, is also clearly essential, as is the final classification step (since the point of the exercise is to positively identify candidate volcanoes in a given image). A high false alarm rate at this point is acceptable assuming the classification component can subsequently discriminate between true detections and false alarms.

Treatment of the three subproblems independently is suboptimal in general, Nonetheless wc treat all three problems independently for the pragmatic reason that one can estimate the parameters of each component in a relatively efficient manner whereas joint estimation of the parameters of the detection, feature extraction, at id classification methods would likely be both comput at ionally impractical and require much larger training set sizes than we have available for this problem. We note in passing that the decomposition of statistical pattern recognition problems into a 2-step process, feature extraction followed by classification, has long been recognized as a necessary evil in most practical pattern recognition problems [21],

G. 1 Detection of Candidate Volcano Regions

The detection component is designed to take an image as input and produce as output a list of candidate volcano locations. A reasonable approach to detection is to usc a matched filter, i.e., a linear filter that matches the signal one is trying to find. For detecting a known signal in white Gaussian noise, the matched filtering approach is optimal, Of course, the volcano problem does not satisfy these underlying assumptions. The set of observed volcanoes cannot be described as a known signal plus white noise, because there is structured variability due to size, type of volcano, surface roughness, etc. Likewise, the clutter background cannot be properly modeled as white noise. Nevertheless, we have empirically found that the following modified matched filtering approach works well.

Let vi denote a k x k pixel region around the i-th training volcano. Each region can be normalized with respect to the local DC level and contrast as follows:

$$\tilde{\mathbf{v}}_i = \frac{\mathbf{v}_i - \mu_i.\mathbf{1}}{\sigma_i} \qquad (1)$$

where μ_i is the mean of the pixels in $\mathbf{v}_i$ and σ_i is their standard deviation. We construct a modified matched filter f by averaging the normalized volcano examples in the training data.

Applying the matched filter to an image involves computing the normalized cross-correlation of f with each k x k image patch. The cross-correlation can be computed efficiently using separable kernel methods to approximate the 2-D kernel f as a sum of 1-D outer products [22].

High response values indicate that there is strong correlation between the filter and the image patch. A typical filter and response image arc shown in Figure ??. Candidate volcano locations arc placed where the matched fi 1 tcr response exceeds a threshold that is determined from training images. Any threshold crossings within a prescribed distance from each other are attributed to the same object and grouped together: the default, distance for the algorithm is 4 pixels.

Detection results on a typical image arc shown in Figure 5. The detected regions of interest are displayed as boxes overlaid on image, while the reference label locations (according to a geologist) arc shown as circles. Although there arc quite a few false alarms, recall that the goal of the matched filter detector is to achieve a low-miss rate while reducing the amount of data to be processed by later stages. Typically the detector is successful in detecting all the volcanoes from Categories 1 and 2, but misses some from Categories 3 and 4.

Although the matched filter can be justified based on empirical results, we also offer the following arguments. First, the k x k windowing eliminates some of the inherent volcano variability, especially that due to scale. Focusing on the central area of the volcano makes the volcano detection problem more like that of finding a known signal since there tends to be less variability in volcano appearance a t the center naturally, the disadvantage is that potentially valuable information outside the

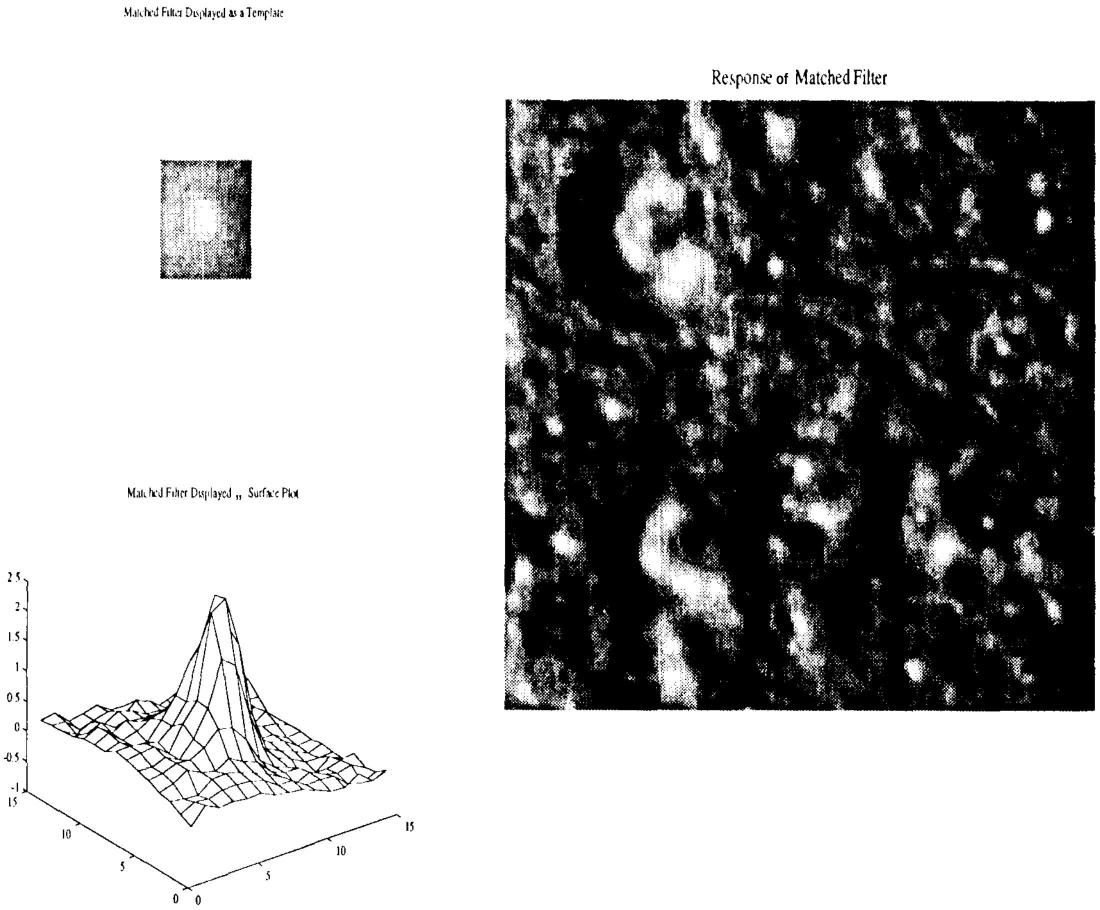

Figure 4: Left: The matched filter contains ninny of the features that planetary geologists report using when manually locating volcanoes. In particular, the matched filter encodes a bright central spot corresponding to a volcanic summit pit and the left-to-right bright-dark shading. Right: Response of tile matched filter on the area shown in Figure 1. Bright points indicate a strong match these will be selected as candidate locations,

k x k window is ignored. Second, normalizing each image patch with respect to the DC level and contrast causes non-descript clutter areas to resemble zero-lnean, white noise. Hence the filter f should be suitable for discriminating these non-descript regions from volcanoes --- the primary purpose of the detector. Of course, in regions where the clutter has features such as graben (narrow ridge-like features on the surface of Venus), the matched filter is not ideal and will produce more false alarms.

6.2 Feature extraction

Since the regions of interest (ROIs) identified by the detector arc embedded in a high dimensional pixel space, the set of possible features is immense. In the results reported here we restrict our search to the family of features defined by linear combinations of the ROI pixel values. This strategy is equivalent to projecting the n-dimensional pixel space onto a q-dimensional subspace (feature space) .

The method of principal components has been used extensively in statistics, signal processing (Karhunen-Loeve transform), and pattern recognition (Turk and Pentland [23]), The problem

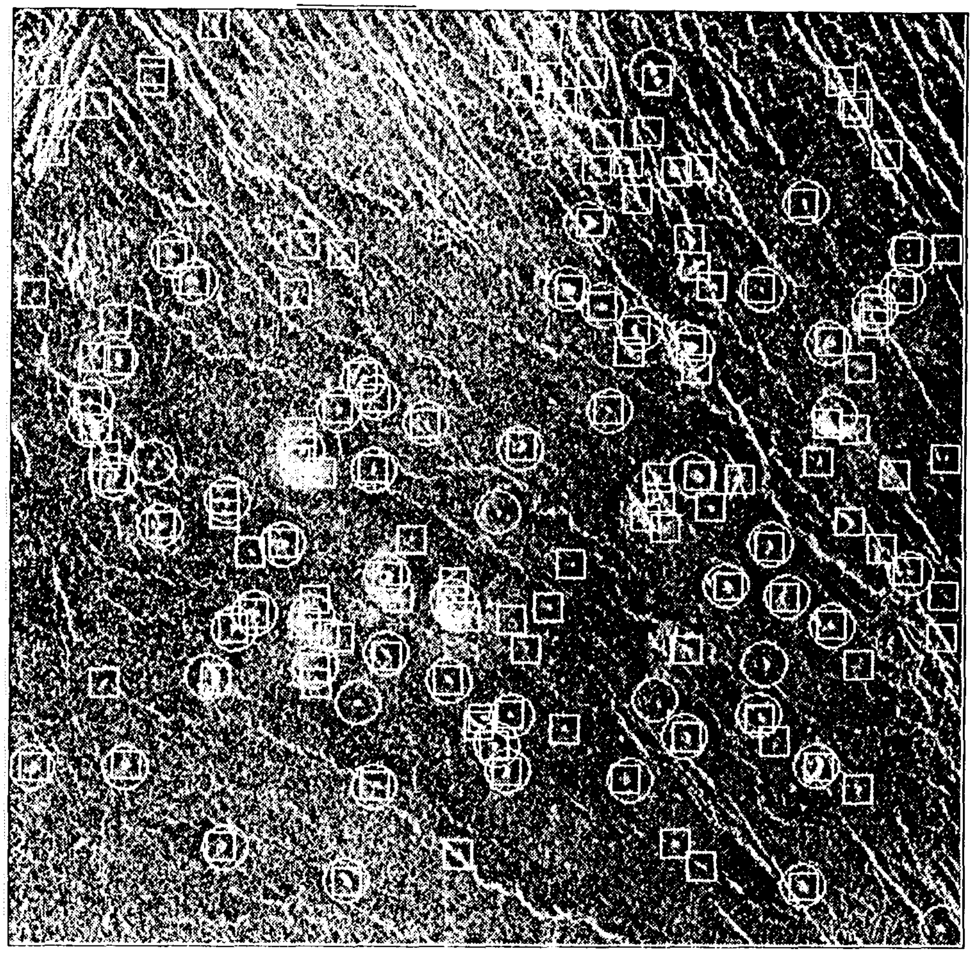

Figure 5: The output of the matched filter detector on a typical image. Circles show the consensus ground truth volcano locations, while boxes show the candidate regions selected by the matched filter detector. Thus, circles with boxes arc detected volcanoes. Circles without boxes are missed volcanoes and boxes without Cil'tics are false alarms. Since the matched filter acts as prescreening for other stages, the cost of a missed volcano is high compared to a false alarm.

formulation is to find a q-dimensional subspace such that the projected data is closest in L_2 norm to the original data, The subspace we seek is spanned by the highest-eigenvalue eigenvectors of the data covariance matrix. Although the full covariance matrix cannot be computed reliably from the number of examples we typically have available, the approximate basis vectors can be computed using the singular value decomposition (SVD) *as* described below,

Each normalized training volcano is reshaped into a vector and placed as a column in an $n \times m$ matrix X, where n is the number of pixels in an ROI and m is the number of training ROIS which contain volcanoes. With tile SVD, X can be factored as follows:

$$X = USV^T \tag{2}$$

For notational convenience, we will assume m is less than n. Then in Equation 2, U is an n x m matrix such that $U^TU = I_{m \times m}$, S is m x m and diagonal with the elements on the diagonal (the singular values) in descending order, and V is m x m with $V^TV = VV^T = I_{m \times m}$. Notice that any column of X (equivalently, any ROI) can be written exactly as a linear combination of the columns of U. Furthermore, if the singular values decay quickly enough, then the columns of X can be closely approximated using linear combinations of only the first few columns of U. That is, the first few columns of U *serve as an* approximate basis for the entire set of examples in X.

The best q-dimensional subspace on which to project is the one spanned by the first q columns of U. The columns of U are shown in Figure 6-b reshaped into ROIs; wc refer to these as features or templates. Notice t hat the first t en templates exhibit struct ure while the remainder appear very random. This suggests projecting onto a subspace of dimension $\leq$ 10. The singular value decay shown in Figure 6-c also indicates that 6 to 10 features encode most of the information in the examples.

Having determined q, we project an ROI into feature space as follows:

$$\mathbf{y} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_q \end{bmatrix}^T \mathbf{x} \qquad (3)$$

where x is the ROI reshaped as an n-dimensional vector of pixels, $\mathbf{u}_i$ is the i-th column of U, and y is the q-dimensional vector of measured features. These feature vectors will serve as input to the classification algorithm.

6.3 Classification

Up to this point in the processing, we have eschewed using counter-examples for training (the detection filter and PCA features were determined solely based on volcanoes). The classifier could also be designed this way, but as shown in [21] such an algorithm is subject to considerable error even in relatively low dimensions because the location of the "other" distribution is unknown. To overcome th is problem, we have experimented with various supervised two-clam methods including quadratic classifiers, decision trees, nearest neighbors, kernel density estimation, and feedforward neural network models. Very similar results were obtained with all of these methods, hence, the quadratic classifier is favored duc to its simplicity,

The quadratic classifier is optimal if the class-conditional probability densities of the feature vector y arc multivariate Gaussian. Assuming y has the postulated claw-conditional densities, the posterior probability that an ROI is a volcano can be estimated using Bayes rule:

$$p(v|\mathbf{y}) = \frac{p(\mathbf{y}|v)p(v)}{p(\mathbf{y}|v)p(v) + p(\mathbf{y}|\bar{v})p(\bar{v})} \qquad (4)$$

where $p(v)$ and $p(\bar{v})$ arc the respective prior probabilities, and

$$\begin{aligned} p(\mathbf{y}|v) &= N(\mathbf{y}, \mu_v, \Sigma_v) \\ p(\mathbf{y}|\bar{v}) &= N(\mathbf{y}, \mu_{\bar{v}}, \Sigma_{\bar{v}}) \end{aligned} \qquad (5)$$

with the notation $N(\mathbf{y}, \mu, \Sigma)$ denoting the multivariate Gaussian density with mean μ and covariance Σ. Onc can show that thresholding the posterior probability in Equation 4 is equivalent to partitioning the feature space with a quadratic hypersurface.

7 Experimental Comparison of Human and Algorithm Detection Performance

In this section, we present the experimental results' obtained using our algorithm to locate small volcanoes in Magellan SAR imagery. The performance of the algorithm in the volcano-location task is compared to the performance of individual geologists, relative to a set of reference labels.

7.1 ROC Methodologies for Performance Evaluation

In its simplest form the ROC plots detections (the system/human detects an object at a location where a volcano exists according to the reference list) versus false alarms (the system/human detects an object where no volcano exists according to the reference list). For a detection system

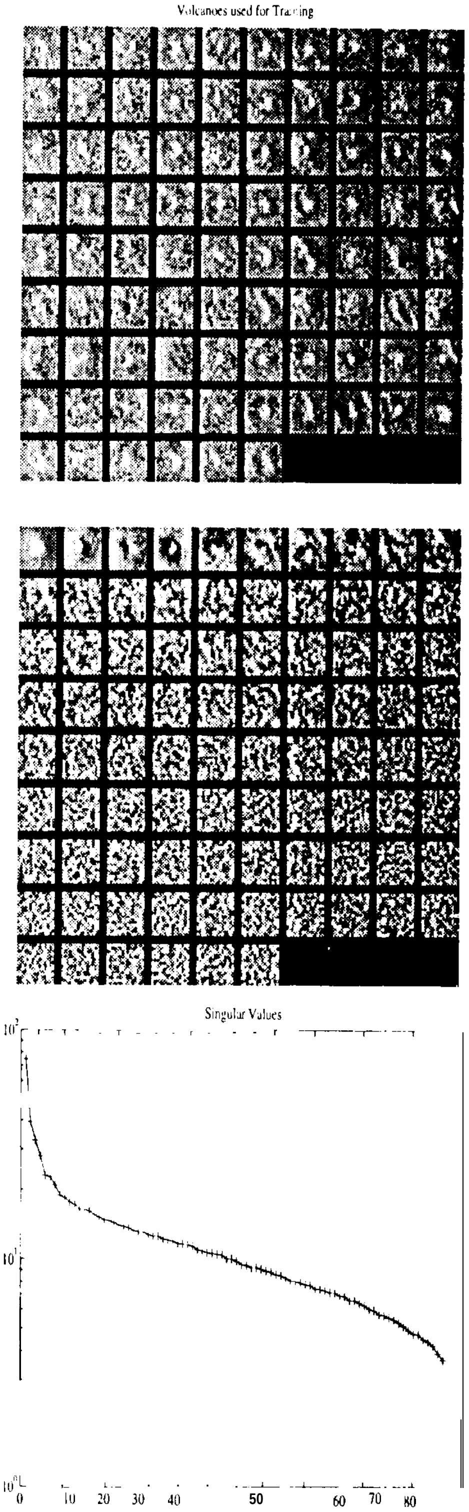

Figure 6: (a) The collection of volcanoes used for feature synthesis. (b) The principal components derived from the examples. (c) The singular values indicate the importance of each of the features for representing the examples.

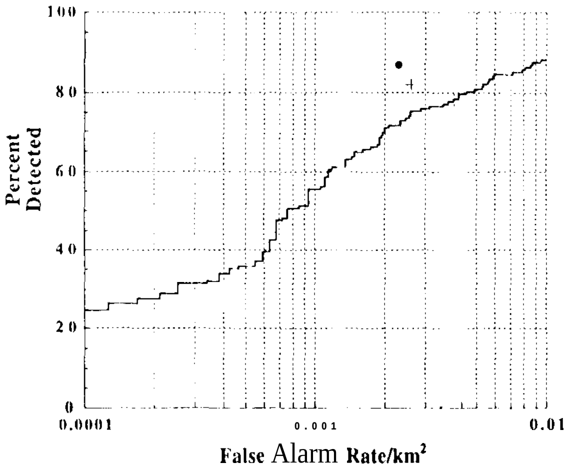

Figure 7: FROC comparing the default algorithm and the geologists based on cross-validation of 4 homogeneous images, using the consensus labeling of the 2 geologists as reference.

which produces posterior probabilities (such as the Gaussian classifier of the previous section), a sequence of (detection/false-alarm) points can be plotted from the test data by varying the posterior probability threshold at which a test region of interest is classified as a volcano.

Ordinarily the maximum number of possible false alarms is fixed in the ROC context and thus the x-axis corresponds to the probability of detecting a false alarm. In the detection of objects (such as volcanoes) in a set of images, the maximum number of potential false alarms (all pixels not close to the object) is not well-defined. A practical alternative to the standard ROC (plotting detection rate versus false alarm rate) for cases such as this is the so-called "free-response" ROC (FROC) [24] which plots the detection rate against the false alarm rate per *unit area.* In the experimental results described in this paper we use the FROC methodology where the x-axis corresponds to false alarms per square km. It is important to note that the FROC can not be analyzed in the same manner as the ROC: for example while the area under an ROC curve can be interpreted as a measure of the quality of the detector there does not exist an analogous quantity for the FROC.

7.2 Experimental Methodology

The experiments described below were conducted using cross-validation: the algorithm was trained on training images and evaluated on a disjoint set of unseen images, and the process repeated over all such training/test pairs of sets. The exact data (names of the training and test images as found on the publicly available CD-ROMS) for each experiment are listed in Appendix 1.

Training consists of a 3-step process based on the training images:

1. Construct the detection filter using the volcanoes in the training images (according to the reference labels for the training images).

2. Determine the principal component directions from the volcanoes in the training images detected in step 1.

3. Estimate the parameters of the Gaussian classifier, using the features from step 2 evaluated on all of the local regions detected in the training data by step 1.

The default settings for algorithm parameters are described in Appendix 2. In general, algorithm detection performance has been found to be relatively insensitive to the exact values of these parameters: experimental results on parameter sensitivity are reported in Section 7.6.

7.3 Experimental Results on A Small Set of Homogeneous Images

In previous work we have reported the results of preliminary experiments using cross-validation on four images that contained 163 small volcanoes and covered a 150km x 150km area of the planet [1 7]. The FROC result is shown in Figure 7. All results were scored relative to the geologists' consensus labeling with confidence categories 1-4 treated as true volcanoes. These 4 images were located rather close together, a factor whose importance will become important as we proceed, For these 4 images the detection performance of the algoi ithm is quite close to that of the geoogists.

7.4 Experimental Results on a Large Set of Homogeneous Images

The small set of' 4 images described above arc part of a 7 x 8 rectangle of 56 images. Of these images, 14 are virtually completely blank clue to a gap in the Magellan data acquisition process, leaving 38 other (42 minus the 4) images to work with. Details on which images were used in the experiments can be found in Appendix 2. The 38 images contained about 480 volcanoes in total and for each training/test partition there were roughly 400 volcanoes in the training image set and 80 in the test set. The performance of the end-to-end algorithm using the default parameters on 6 different partitions of this 38 image data set is shown in Figure 8 using the labels of geologist A as the reference, and in Figure 9 using the labels of geologist B. The solid curve is the measured FROC on the test set for the algorithm. The solid circular symbol in each plot is the performance of geologist, A (B) relative to the labeling of B (A). The "+" symbol is the performance of one of the authors (MCI]). We note that the two geologists are operating at different parts of the FROC curve (comparing the plots of Figure 8 and Figure 9): geologist B is relatively conservative relative to geologist A. The ~lon-expert, MCB, is quite close in performance to geologist A (Figure 9), tending to have a somewhat higher false alarm rate and slightly higher detection rates. The performance of tile algorithm is reasonable but not as accurate as the humans. In Figure 9 the algorithm is between 10 to 50 % below the detection accuracy of the humans at a fixed false alarm rate. Using geologist A as reference, the algorithm performs somewhat better, being between 5 to 20% less accurate in terms of detect ion performance (Figure 8).

For a particular training/test partition we evaluated the performance of the matched filter alone as a det ector and compared its performance to the matched filter combined with the feature ext ract ion and Gaussian classifier (the default algorithm). The results are shown in Figure 10 in the same FROC format as before, The detector has a free parameter (a threshold) that controls its aggress iveness in declaring volcanoes, i.e., the trade-off between misses and false alarms. Varying this parameter generates an F] tOC curve for the detector alone (wit bout the Gaussian classifier). Observe that the combination of matched filter and classification yields better performance than using only a matched filter (use of a matched filter alone was proposed in [7]).

7.5 Experimental Results on Inhomogeneous Images

In this experiment 36 images were selected from random locations on the planet. These 36 images contained significantly greater variety in shape, noisiness, and size than the sets of 4 and 38 used in the earlier experiments. There were about 670 volcanoes in total in the 36 images, with about 500 in the training set and 170 in the test set for each partition. Figure 11 shows the FROC performance from 4 different partitions of the data into 27 training images and 9 test images (details in Appendix 2). Clearly the system is performing worse than on the more homogeneous image sets. For example at the 0.001 false alarm rate/km^2 the detection performance is in the 20-40% range whereas for the 38 homogeneous images the detection rates were consistently in the 50- 65% range. For this data set the reference labels arc consensus labels (where geologists A and B jointly labeled the images): for the few images (of the 36) where we have individual labels in addition to the consensus labels, the geologist's detection performance appears to be in the same general region as it was for the homogeneous images. Thus, one can conclude that the volcano detection approach does not handle image inhomogeneity as well as human experts. This is to be expected since both the principal components and Gaussian classification models are essentially based on the assumption that the volcano population can be described in a unimodal fashion in terms of pixel appearance and size, whereas with inhomogeneous images there may be multiple sub-classes present. '1'1 ius, more complex models are likely to be necessary to handle the inhomogeneous image case.

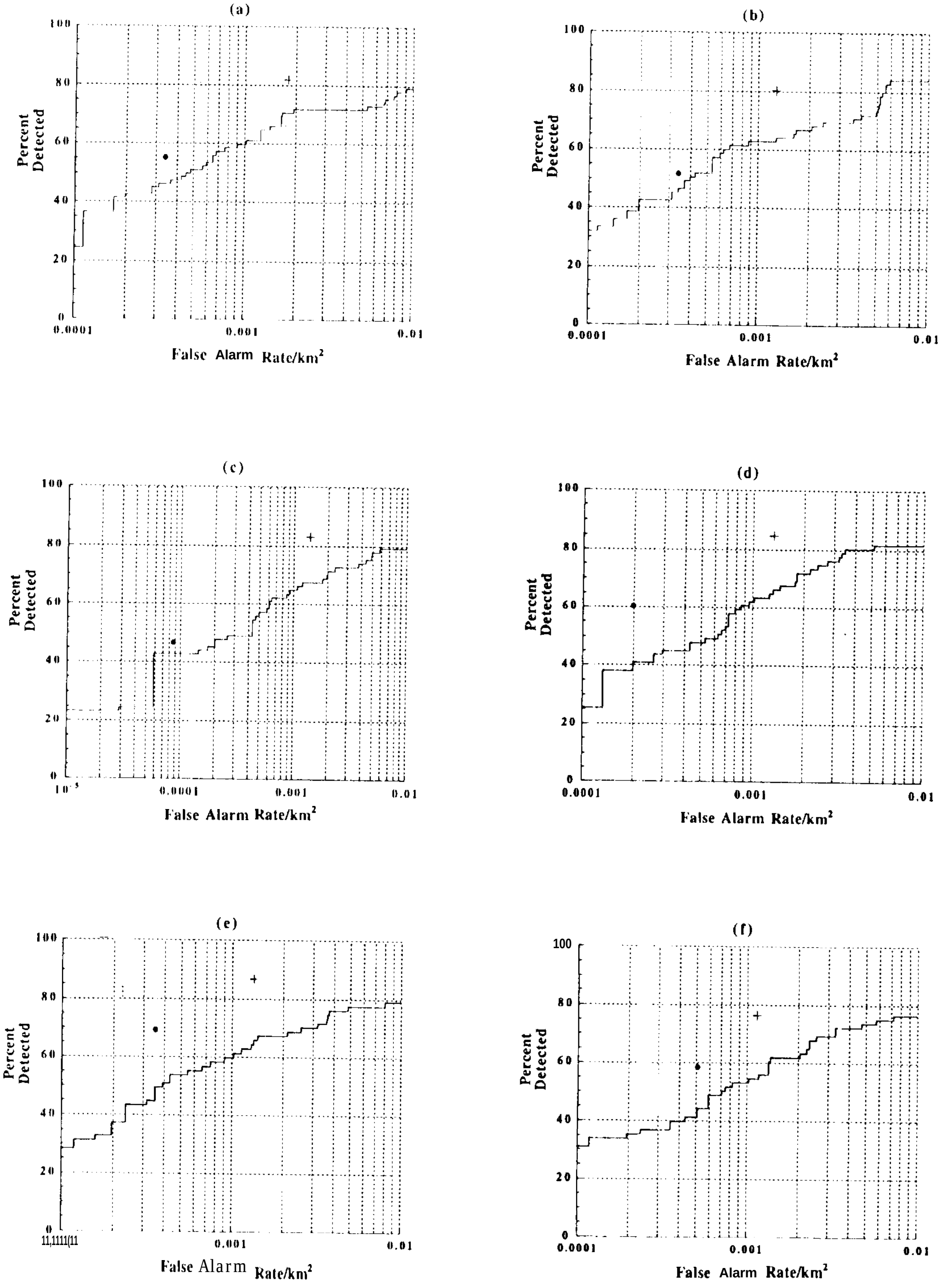

Figure 8: FROC comparing the default algorithm and geologist B based on training on 32 images and testing on 6, 6 different partitions, using the labeling of geologist A as reference.

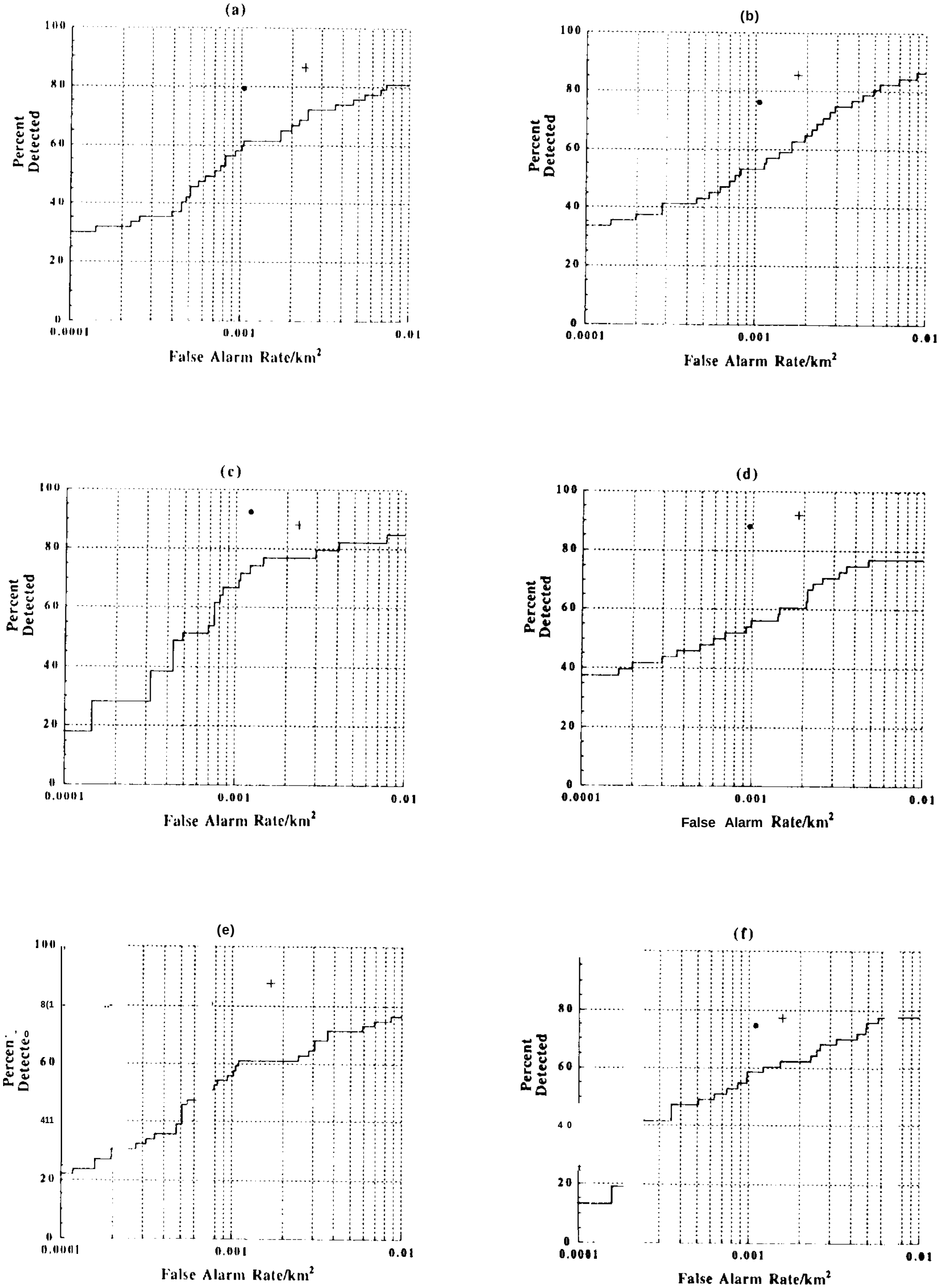

Figure 9: FROC comparing the default algorithm and geologist A based on training on 32 images and testing on 6, 6 different partitions, using the labeling of geologist B as reference.

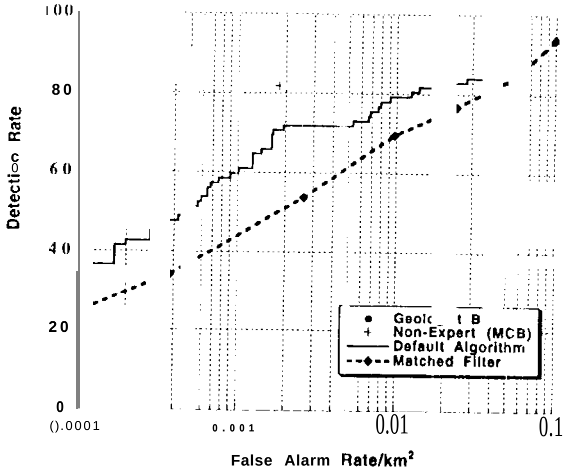

Figure 10: FROC comparing the matched filter detector, the overall default algorithm and geologist B, training on 32 images and testing on 6 (partition (a)), using the labeling of geologist A as reference.

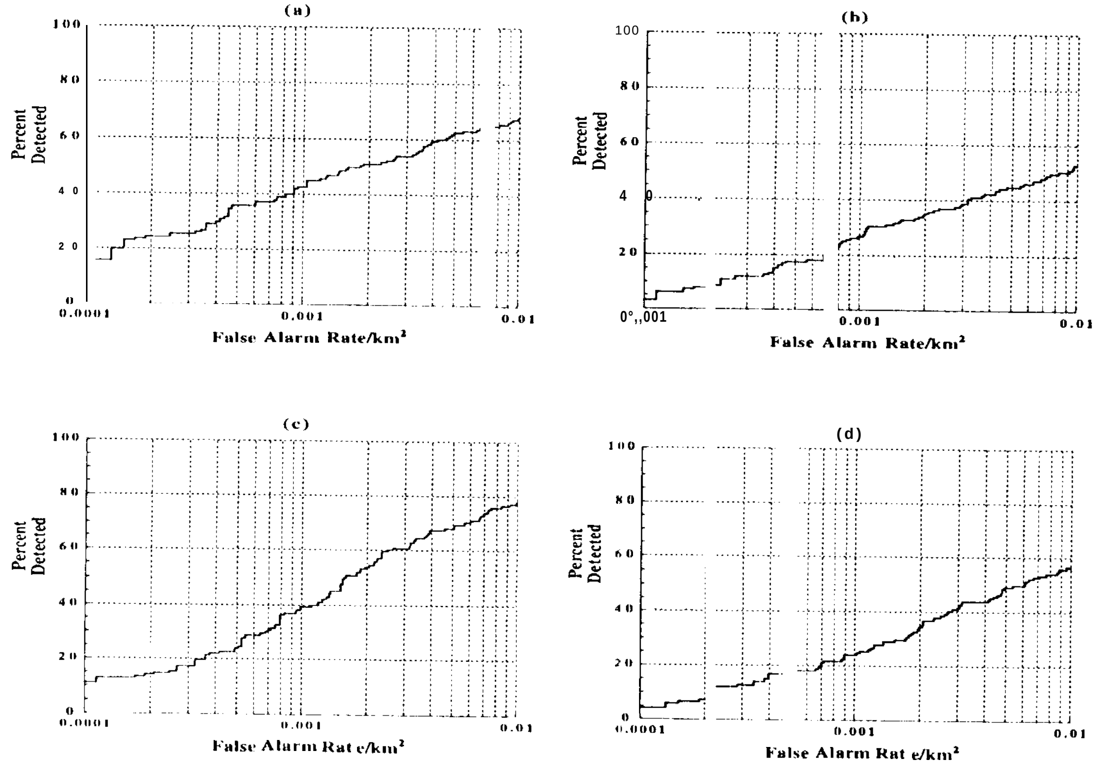

Figure 11: FROC for the default algorithm based on training on 27 images and testing on 9, 4 different partitions, using tile consensus labeling for the 36 images as reference.

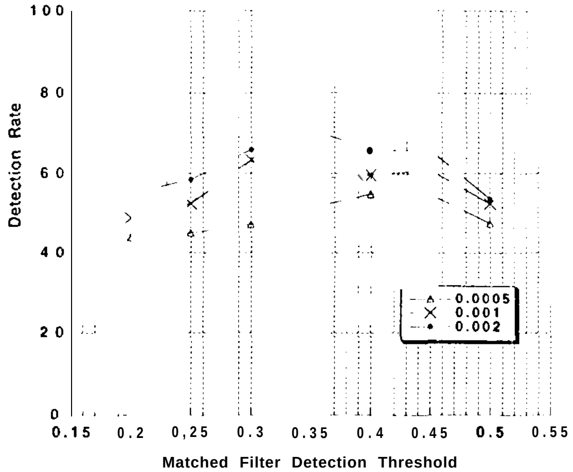

Figure 12: Overall detection performance as a function of the matched filter detection threshold evaluated at three different fixed false alarm rates (0,0005, 0.001, ().002).

7.6 Algorithm Sensitivity to Parameter Values

How dots tile algorithm's detection performance vary if the various parameters are changed? Ideally onc would like a relatively stable operating range so that the algorithm is not over-sensitive to the exact values of the parameters. We do not include in this paper any sensitivity results on the size of the detection or principal component windows or the detection clustering threshold or radius of detection parameters: t hc defaul t values for these parameters were chosen based on knowledge of tile typical volcano sizes and informal experimental results have shown the algorithm to be relatively insensitive to the exact values.

Of greater interest is the algorithm sensitivity to the threshold parameter used at the matched filter detection stage and the number of principal components used as features for classification. In both c ases below the detection rate is estimated as a function of the parameter of interest for three different false alarm rates. The three rates chosen were 0,0005, 0.001, and 0.002 false alarms per km^2 which roughl y correspond to t he range of' opcrat ing points used by humans. For t hc purposes of illustration the results below arc for onc particular train/test combination (combination (a) from the previous section, train on 32 images and test on 6). However similar qualitative results have been observed across a variety of training and test image sets for both of these parameters, In both cases only the parameter of interest is varied and the other parameters are held at their default values: investigations into the multivariate performance dependence on multiple parameters was not feasible given t he amount of data available for these experiments.

7.6.1 Sensitivity to Matched Filter Detection Threshold Parameter

Note that the operating detection rate from the matched filter is necessarily an upper bound on the detection rate of classification algorithm since volcanoes missed at the matched filter stage are missed forever. Thus, it would appear that onc would prefer to be at the highest possible matched filter detect ion rate. However, it is not clear whether a somewhat lower detection rate might be better in an overall sense since there may be a much lower proportion of false alarms for the classifier to deal with in the feature space.

In Figure 12 the detection rate of the classifier is plotted as a function of the matched filter threshold, for the t hree different fixed false alarm rates. In the 0.3 to 0.45 range of operation, performance appears somewhat sensitive to the exact value of the threshold, but nonetheless this appears to the optimal operating range. If the threshold is below 0.3, the detection rate tails off because although the det ector is detecting more true volcanoes this is traded-off with the fact that it is detect ing orders of magnitude more false alarms. The increase in false alarms has much more of an effect in terms of final classification than the relatively small increase in true detections. Above 0.45, even though there arc fewer false alarms detected, there are too fcw volcanoes detected by

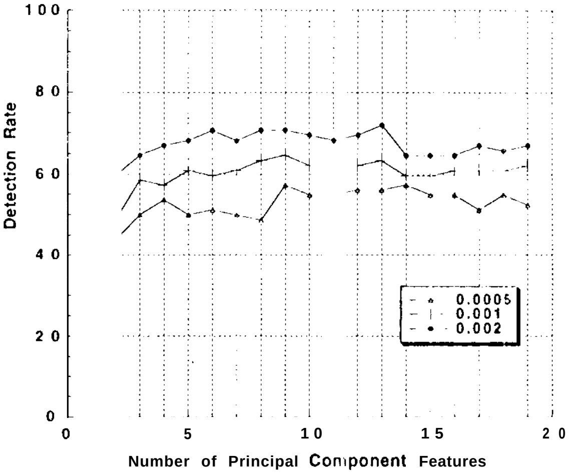

Figure 13: Overall detection performance as a function of the number of principal component features used for classification evaluated at three different fixed false alarm rates (0.0005, 0.001, 0,002).

the matched filter, and thus the overall detection performance of the algorithm is upper-bounded.

7.6.2 Sensitivity to Number of Principal Components

The default algorithm uses the first 6 principal component as features: the number 6 was chosen as a trade-off between retaining a certain number of the more informative principal components and keeping the feature space dimensionality low. Figure 13 shows how the detection rate varies for different numbers of principal components, for three different fixed false alarm rates. Apart from a marked decrease 'in detection rates mice the number of features goes below 4, detection performance is relat ively constant over the rest of the range.

8 Lessons Learned with Implications for Digital Image Library Applications

8.1 The Feasibility of Lemming from Pixel-Level Descriptions

While it is appealing to consider a user pointing to a few examples of interest and having the system then learn a detection model, such a "bottom-up" approach based on learning alone may not scale well to difficult problems. As pointed out in the paper by Geman, Bienenstock and Doursat [25], an algorithm which learns from pixels alone is operating in such a high-dimensional space that statistical estimation theory predicts that prohibitively large amounts of training data arc required to reduce the variance of the estimates, i.e., to construct an accurate detection model from data. The authors conclude that the use of appropriate prior knowledge embedded in the model is the only practical way to circumvent this problem. In effect, the principal components methodology described earlier embodies a limited foi m of prior knowledge in the form of a belief that the volcano population can be described as linear combinations of a few "basis" volcanoes. However, the performance on the inhomogeneous image sets shows that this particular prior bias may be inappropriate for the more general volcano detection problem where sub-classes of volcanoes may be present.

On the other hand it is difficult to see how a purely model-based, non-adaptive approach could work for a problem of this nature. The Geologists provide descriptions of the visual cucs they use in detecting volcanoes such as "bright-dark pairs," "circular outline, " etc. Translating these high-level descriptions into pixel-level constraints is virtually impossible since standard shape extraction methods 1)ased on edge and segment informat ion arc not well- matched to noisy natural images such as the volcano data.

‘]’bus, a middle ground between the model-based and learning approaches appears likely to be the most practical avenue for building aut omated recognition syst ems for datasets of this nature. In problems where the feature definition problem can be solved Ul)f’milt by the user, the recognition rates are typically much higher (e.g., [26]). For more general and difficult problems, an interactive capability for feature definition involving both the expert and the image data is required.

8.2 Subjective Elements in Image Analysis

One of the primary lessons learned from this project is the importance of the subjective human element in model training and performance evaluation. Image analysis by humans is a subjective process. Thus, for many digital library tasks where t he quality of the result is subject, to human interpretation it is critical that the sill.),jcctivc aspect of the! process is taken into account. For the volcano project we have primarily adopted the simple approach of fixing one expert’s (or set of experts) subjective estimate as ground truth for each experiment and then evaluating all other estimates against this reference. However this is sul)opt imal in t he sense that in the worst case one expert’s opinion might be no better than random and performance estimates using that expert as reference should receive lower weight. We have i ivestigat ed soi ne probabilistic techniques for modelling multiple expert opinions [18] and there is a significant body of work in the statistical and biomedical literatures on this topic [27, 28, 29]. However, since little or none of this work concerns rat ing expert opii ion based on *visual* stimuli there is clear] y room for much more work on this topic given its fundamental imp ortance in problems i ivolving detect ion and cataloging subject to human review,

8.3 Invariance issues

Despite its intuitive appeal, t 1 iere are a number of arguments against using the sort of simple template-based approach for detection and cataloging we have described in this paper. Most notably, the proposed method is not invariant, with ı espect to translation, rotation, scaling, and direction of illumination. A certain (hopefully small) number of templates will be required in order to represent t he inherent variability of an object; any additional variability due to spatial shifting, rotation, scaling, or noise will dramatically increase the number of templates required to encode the object. For example the performance of the system was significantly worse on less homogeneous sets of training and test images (Figure 11). Thus, the ı emplat c-based approach may not be feasible unless appropriate normalization steps are taken prior to feature learning. These invariance issues need to be resolved in order to develop a general systei i; however, for the volcano problem they are not so critical since (1) the detection step effectively “centers” the volcanoes well, (2) the volcanoes have significant rotational symmetry, (3) the central ai ea of the volcanoes (on which the templates are based) are relatively insensitive to overall scale, and (4) the direction of illumination is known and relatively constant.

Note that for general tasks of cataloging objects in digital libraries consisting of uniformly gathered ant] processed data (e.g., a fixed remote sensing platform, documents scanned from a single source, or records of patients treated in some uniform manner) a certain degree of invariance can be expected. The volcano problem is one such example.

8.4 The Need for An Adaptive Approach

The volcano det ection and cataloging problem is a good example of a situation that is becoming all too common in many fields, spanning science data analysis, medical image analysis, commercial graphic arts, surveillance, and so forth. The volumes of data are so large that comprehensive manual analysis and search is not possible. Since most users are not programmers or experts in pattern recognit ion, an adaptive approach based on learning from examples is gradually becoming a *necessity* in some settings. Work on developing robust algorithi ns to address such needs is very much needed.

no hyphens

9 Conclusion

This paper discussed the general problem of translating a large image dataset into a catalog of objects of interest, in particular, the problem of detecting and cataloging small volcanoes on the surface of Venus. Scientific users are often not interested in the image data *per* se but in derived products such as catalogs and libraries of objects of interest: these catalogs form the basis for quantitative scientific analysis. A trainable detection system for automatically generating volcano catalogs was discussed, Experimental results showed that the system is approaching human performance on homogeneous sets of images but performs poorly on inhomogeneous image sets. Combining prior information with data and modeling subjective human opinion were both identified and discussed as key issues in problems of this nature.

This paper aims to provide an example of an important large-scale application in the area of aiding humans in the analysis of a large digital library, A secondary aim is to emphasize the need for a natural interface between humans and digital libraries: one where the user can interact directly with the library contents without the need for a programmer (or a team of programmers) in the loop to produce customized pattern recognizers for each cataloging and recognition task. A learning-from-examples approach could provide the basis for such a practical and natural interface for certain classes of lar?,c-scale digital image data sets.

Acknowledgements

The research described in this report has been carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Support was provided primarily by NASA Office of Space Access and Technology (SAT - Code X), a J] 'I. DDF award, and NSF research initiation grant IRI 9211651. We would like to thank geologists Jayne Aubele and Larry Crumpler of Brown University for their assistance in labeling and analyzing the Magellan data. We would also like to thank Maureen Burl and Joe Roden for help in developing the software and user-i)tcrfaces.

References

[1] *Magellan at Venus: Special Issue of the Journal of Geophysical Research,* American Geophysical Union, 1992.

[2] *NSSDC News*, vol. 10, no.1, Spring 1994, available from *request@nssdc.gsfc.nasa.gov.*

[3] Guest, J. E. et al. 1992. Small volcanic edifices and volcanism in the plains of Venus. *Journal of Geophysical Research*, vol.97, no.E10, pp.15949-66., October 25, 1992.

[4] J. A. Richards, *Remote Sensing* /or *Digital image Analysis,* Springer-Verlag, Berlin, 1986.

[5] A. M. Cross, "Detection of circular geological features using the Hough transform, " *Int. J. Remote Sensing,* 9, no.9, *1519-1528,* 1988.

[G] J. Skingley and A. J. Rye, "The Hough transform applied to SAR images for thin line detection," *Pattern Recognition Letters,* 6, *61-67,* June 1987.

[7] C. R. Wiles and M. R. B. Forshaw, "Recognition of volcanoes using correlation methods," *Image and Vision Computing,* vol.11, no.4, pp.188-196, 1993.

[8] R. S. Saunders et al., Magellan mission summary, *Journal of Geophysical Research,* vol.97, no. E8, pp.13067- 13090, August 25, 1992.

[9] *Science*, special issue on Magellan data, April 12, 1991.

[10] *1'. Cat t* ermolc, *Venus: The Geological Story*, Baltimore, MD: Johns Hopkins University Press, 1994.

[11] J. C. Aubele and E. N. Slyuta, "Small domes on Venus: characteristics and origins," in *Earth, Moon and Planets*, *50/51 , 493-532,* 1990.

[12] G. H. Pettengill et al., "Magellan: radar performance and product s," Science, vol.252, 260- 265, 12 April 1991.

[13] 1'. C. Bunch, J. F. Hamilton, G. K. Sanderson and A. H. Simmons, "A P'rcc-Response approach to the measurement and characterization of radiogral phic-observer performance," *J. Appl. Photo. Eng.*, vol.4, no. 4, pp.166-171, 1978.

[14] M. S. Chesters, "Human visual perception and ROC methodology in medical imaging," *Phys. Med. Biol.*, vol.37, Ilo.7, pp.1433-1476, 1992.

[15] P. Smyth, 'Learning with probabilistic supervision, ' in *Computational Learning Theory and* Natural *Learning Systems 3*, *'1'.* Petsche, S. Hanson, and J. Shavlik, Cambridge, MA: MIT Press, pp.163–182, 1995.

[16] I). Kahneman, I'. Slovic, and A. Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases,* Cambridge University Press, 1982.

[17] M. C. Burl, U. M., Fayyad, P. Perona, P. Smyth, and M. P. Burl, "Aut omating the hunt for volcanoes on Venus," in *Proceedings of the 1994 Computer Vision and Pattern Recognition Conference, C VPR-94,* Los Alamitos, CA: IEEE Computer Society Press, pp.302–309, 1994.

[18] P. Smyth, M. C. Burl, U. M. Fayyad, P. Perona, 'Knowledge discovery in large image databases: dealing with uncertainties in ground truth," in *Advances* in *Knowledge Discovery and Data Mining,* U. M. Fayyad, G. Piatetsky-Shapiro, I'. Smyth, R. Uthurasamy (eds.), AAAI/MIT Press, in press.

[19] M. C. Burl, U. M. Fayyad, P. Perona, and P. Smyth, "Automated analysis of radar imagery of Venus: handling lack of ground truth," in *Proceedings of the IEEE Conference on Image* Processing, Piscataway, NJ: IEEE Press, vol.III, pp.236–240, 1994.

[20] P. Baldi, personal communication, 1994.

[21] K. Fukunaga, *Statistical Pattern Recognition,* 2nd cd,, Academic Press, 1990.

[22] S. Treit el and J. Shanks, "The design of multistage seperable planar filters," *IEEE Trans* Geoscience *Electron,* GE-9(1):10–27, 1971.

[23] M. Turk and A. Pentland. "Eigenfaces for recognition." *J. of Cognitive Neurosci.*, 3:71–86, *1991.*

[24] *D. 1'.* Chakraborty and L. H. L. Winter, "Free-Response methodology: alternate analysis and a new observer-per formance experiment," *Radiology, 174,* 873–881, *1990.*

[25] *S.* Geman, E. Bienenstock and R. Doursat, 'Neural networks and the bias/variance dilemma,' *Neural Computation, 4,* pp.1–58, 1992.

[26] U. M. Fayyad, I'. Smyth, M. C. Burl, ancl P. Perona, P., "A learning approach to object recognition: applications in science image database exploration and analysis," in *Early Visual Learning, S.* Nayar and T. Poggio (eds.), in press.

[27] J. S. Uebersax, "Statistical modeling of expert ratings on medical treatment appropriateness," *J.* Amer. *Statist. Assoc.*, vol.88, no.422, pp.421- *427, 1993.*

[28] A. Agresti, "Modelling patterns of agreement and disagreement," *Statistical Methods in Medical Research*, vol.1, pp.201 218, 1992.

[29] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics,* vol.28, no.1, *pp.20 -28, 1979.*

Appendix 1: Obtaining the Magellan Dataset and Lists of Images Used in the Experiments

How to obtain the Magellan Images and Labels

Note to the Referees ant] Editor: Wc are in the process of constructing a WWW page which will allow direct access to both the image data and label data described in this paper. This WWW page will be accessible via `http: //www-aig.jpl.nasa. gov/mls/mgn-sar.`

Lists of which Images were used in each Experiment

Note that location of an image is indicated by a unique directory name (or "product") of the form f$ab$n$xyz$ where ab is the longtitute, n denotes the Northern Hemisphere in this case, and xyz is the latitude. Each product (directory) contains 56 images arranged in a $7x8$ contiguous grid, numbered from top left to bottom right.

The set of 4 images came from directory f30n332 and consisted of : ff05, ff13, ff20, and ff21.

The set, of 38 images consisted of 6 sets of 6 images from directory f30n332 where each image is denoted as ffxy and the xys are organized as follows:

- set (a) *22, 24, 28, 39, 52, 55*
- set (b) 04, *06, 32, 36, 47, 54*
- set (c) 07, 14, 23, 37, 40, 44
- set (d) *03, 15,* 19, *31, 38, 53*
- set (c) 08, 12, 27, 30, 43, 56
- *set* (f) 11, 16, 29, *35, 46, 48*

In addition there were two extra images, *ff45* and ff51 which were not part of any test set and were used in all 6 training sets. The remaining images wei e primarily blank and were not used in the experiments. Each experiment consisted of using one of (a), (b), (c), (cl), (e), and (f) as test set and training on the other images plus ff45 and ff51.

The set of 36 inhomogeneous images were broken down into 4 sets of 9 images, (a)-(d), where for each experiment each of (a)-(d) was denoted the test set and the algorithm was trained on the other 3 sets, The sets were:

- set (a): f40n272-ff34, f05s312-ff33, f30n281-ff19, f50n197-ff26, f25n284-ff37, f40n272-ff24, f10n211-ff54, f10n279-ff38, f75n351-ff47.
- set (b): f50s088-ff36, f10s301 -ff19, f75n237-ff5, f40n286-ff39, f05n284-ff44, f60n302-ff37, f00n279-ff37, f40n244-ff50, f15n129-ff08.
- set (c): f] 0n267-ff01 , f45s012-ff51 , f25s302-ff18, f1 5n020-ff53, f30n332-ff12, f25s302-ff30, f00n318-ff01 , f05s211-ff21, f25n229-ff47.
- set, (()): f] 0n076-ff23, f15n283-ff27, f10s245-ff38, f] 5n283-ff49, f45n188-ff20, f05s290-ff43, f20s257-ff54, f25s1 98-ff54, f55n291-ff45.

For each image the x,y coordinates of the volcanoes as labeled (estimated) by geologist A, B or the consensus of both (depending on which is available for each image) is available from the WWW page mentioned above.

Appendix 2: Default Settings for A lgorithm Parameters

In all of the experiments in the paper, unless otherwise stated, the algorithm parameters were set to default, values whose values were determined manually from experimenting with the set of 4 images described in earlier work *[17]*. In particular,

- In training, all volcanoes are treated equally, i.e., the categories 1-4 are not used to weight the training in any way.
- The window width k_1 for the detector was 30 pixels.
- The threshold value for the detector was 0.35.
- The window width k_2 for the derivation of the SVD decomposition was 15 pixels: these 15 x 1 5 windows were obtained by subsampling the 30 x *30* local regions by a factor of 2.
- The threshold for the detection clustering algorithm was 4 pixels,
- The number of principal components (features) used for clarification was 6.
- The classification method used was a maximum-likelihood Gaussian classifier, with independent full-covariance matrices for each class.
- Let $r_{0.5}$ be half the estimated radius (according to the reference list) of a volcano close to a detected location. A region was declared a detection if the Euclidean distance d between the location of the detection and the location of the volcano on the reference list, was less then $r_{0.5}$, unless $r_{0.5} < 5$ pixels in which case $r_{0.5}$ is replaced by 5, or $r_{0.5} > 15$ pixels in which case $r_{0.5}$ is replaced by 15. Thus, the criterion for a detection was that the detected location be within half the radius of the reference volcano unless the radius is extremely small or ext remely large, Empiricall y it has been found that volcanoes rarely overlap thus effectively eliminating the problem of detecting multiple volcanoes which arc very close together,